Pairwise protein substring alignment with latent semantic analysis and support vector machines to detect remote protein homology

Abstract:

Remote protein homology detection has been widely used as a part of the analysis of protein structure and function. In this study, the good quality of protein feature vectors is the main aspect to detect remote protein homology; as it will assist discriminative classifier model to discriminate all the proteins into homologue or non-homologue members precisely. In order for the protein feature vectors to be characterized as having good quality, the feature vectors must contain high protein structural similarity information and are represented in low dimension which is free from any contaminated data. In this study, the contaminated data which originates from protein dataset was investigated. This contaminated data may prevent remote protein homology detection framework to produce the best representation of high protein structural similarity information in order to detect the homology of proteins. To reduce the contaminated data and extract high protein structural similarity information, some research has been done on the extraction of protein feature vectors and protein similarity. The extraction of protein feature vectors of good quality is believed could assist in getting better result for remote protein homology detection. Where, the good quality of protein feature vectors containing the useful protein similarity information and represent in low dimension will be used to identify protein family precisely by discriminative classifier model. Referring to this factor, a method which combines Protein Substring Scoring (PSS) and Pairwise Protein Substring Alignment (PPSA) from sequence comparison model, chi-square and Singular Value Decomposition (SVD) from generative model, and Support Vector Machine (SVM) as discriminative classifier model is introduced.